`Math 583    HW 10 Fall 2020        Due Friday, Nov. 13.`

Quiz 10 on Friday (no class Wed.) will have problems on this HW. Final: Monday, Dec. 7, 8-10 AM. Problem numbers are from Olive (2020). Do the source commands from homework 4.

7.11. For ridge regression, let $\boldsymbol{A}_n = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{W}$ and $\boldsymbol{B}_n = [\boldsymbol{I}_{p-1} - \lambda_{1,n}(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}]$. Show $\boldsymbol{A}_n - \boldsymbol{B}_n = \boldsymbol{0}$.

8.1. When doing a PI or CI simulation for a nominal $100(1-\delta)\% = 95\%$ interval, there are m runs. For each run, a data set and interval are generated, and for the ith run $Y_i = 1$ if μ or Y_f is in the interval, and $Y_i = 0$, otherwise. Hence the Y_i are iid Bernoulli$(1-\delta_n)$ random variables where $1-\delta_n$ is the true probability (true coverage) that the interval will contain μ or Y_f. The observed coverage (= coverage) in the simulation is $\overline{Y} = \sum_i Y_i/m$. The variance $V(\overline{Y}) = \sigma^2/m$ where $\sigma^2 = (1-\delta_n)\delta_n \approx (1-\delta)\delta \approx (0.95)0.05$ if $\delta_n \approx \delta = 0.05$. Hence

$$SD(\overline{Y}) \approx \sqrt{\frac{0.95(0.05)}{m}}.$$

If the (observed) coverage is within $0.95 \pm kSD(\overline{Y})$ the integer k is near 3, then there is no reason to doubt that the actual coverage $1-\delta_n$ differs from the nominal coverage $1-\delta = 0.95$ if $m \geq 1000$ (and as a crude benchmark, for $m \geq 100$). In the simulation, the length of each interval is computed, and the average length is computed. For intervals with coverage $\geq 0.95 - kSD(\overline{Y})$, intervals with shorter average length are better (have more precision).

a) If $m = 5000$ what is 3 SD($\overline{Y}$), using the above approximation? Your answer should be close to 0.01.

b) If $m = 1000$ what is 3 SD($\overline{Y}$), using the above approximation?

8.2. The smoothing spline simulation compares the PI lengths and coverages of 3 large sample 95% PIs for $Y = m(x) + e$ and a single measurement x. Values for the first PI were denoted by scov and slen, values for 2nd PI were denoted by ocov and olen, and values for third PI by dcov and dlen. The average degrees of freedom of the smoothing spline was recorded as *adf*. The number of runs was 5000. The *len* was the average length of the PI and the *cov* was the observed coverage. One student got the following results shown in Table 4.2.

Table 1: Results for 3 PIs

error		95%	PI	95%	PI	95%	PI	
type	n	slen	olen	dlen	scov	ocov	dcov	adf
5	100	18.028	17.300	18.741	0.9438	0.9382	0.9508	9.017

For the PIs with coverage ≥ 0.94, which PI was the most precise (best)?

8.6. A problem with response and residual plots is that there can be a lot of black in the plot if the sample size n is large (more than a few thousand). A variant of the response plot for the additive error regression model $Y = m(\boldsymbol{x}) + e$ would plot the identity line, the two lines parallel to the identity line corresponding to the Section 7.12 large sample $100(1-\delta)\%$ prediction intervals for Y_f that depends on $\hat{Y}_f$. Then plot points corresponding to training data cases that do not lie in their $100(1-\delta)\%$ PI. We will use $\delta = 0.01$, $n = 100000$, and $p = 8$.

a) Copy and paste the commands for this part into *R*. They make the usual response plot with a lot of black. Do not include the plot in *Word*.

b) Copy and paste the commands for this part into *R*. They make the response plot with the points within the pointwise 99% prediction interval bands omitted. Include this plot in *Word*. For example, left click on the plot and hit the *Ctrl* and *c* keys at the same time to make a copy. Then paste the plot into *Word*, e.g., get into *Word* and hit the *Ctrl* and *v* keys at the same time.

c) The additive error regression model is a 1D regression model. What is the sufficient predictor $= h(\boldsymbol{x})$?

8.7. The Rousseeuw and Leroy (1987, p. 26) Belgian telephone data has response $Y =$ *number of international phone calls* (in tens of millions) made per year in Belgium. The predictor variable $x =$ year (1950-1973). From 1964 to 1969 total number of minutes of calls was recorded instead, and years 1963 and 1970 were also partially effected. Hence there are 6 large outliers and 2 additional cases that have been corrupted.

a) The simple linear regression model is $Y = \alpha + \beta x + e = SP + e$. Copy and paste the *R commands* for this part to make a response plot of $ESP = \hat{Y} = \hat{\alpha} + \hat{\beta} x$ versus Y for this model. Include the plot in *Word*.

b) The additive error GAM is $Y = \alpha + S(x) + e = AP + e$ where S is some unknown function of x. The *R commands* make a response plot of $EAP = \hat{\alpha} + \hat{S}(x)$ versus Y for this model. Include the plot in *Word*.

c) The simple linear regression model is a special case of the additive error GAM with $S(x) = \beta x$. The additive error GAM is a special case of the additive error regression model $Y = m(x) + e$ where $m(x) = \alpha + S(x)$. The response plots for these three models are used in the same way as the response plot for the multiple linear regression model: if the model is good, then the plotted points should cluster about the identity line with no other pattern. Which response plot is better for showing that something is wrong with the model? Explain briefly.